**IN THE UNITED STATES DISTRICT COURT
FOR THE SOUTHERN DISTRICT OF NEW YORK**

HACHETTE BOOK GROUP, INC., HARPERCOLLINS PUBLISHERS LLC, JOHN WILEY & SONS, INC., and PENGUIN RANDOM HOUSE LLC
 	
                Plaintiffs,
 	
    v.
 	
INTERNET ARCHIVE and DOES 1 through 5, inclusive
 	
                Defendants.
	Case No. 1:20-CV-04160-JGK

REBUTTAL EXPERT REPORT OF IMKE REIMERS, PH.D.

ATTORNEYS' EYES ONLY

TABLE OF CONTENTS

i

After completing my initial expert witness report on February 25, 2022, I received a copy of the expert witness report of Jeffrey T. Prince. The Internet Archive asked me to review this expert witness report and to determine whether I should rebut any portion of it. I indeed do find that Professor Prince's report warrants a rebuttal. I therefore wish to provide, below, my rebuttal.

I. PROFESSOR PRINCE CLAIMS I IGNORE THE EFFECT OF THE INTERNET ARCHIVE ON E-BOOK SALES.

1. A main point raised by Dr. Prince is that my analysis excludes the effects of the Internet Archive on e-book sales (e.g., Prince Rpt. ¶¶ 79 & 81, ¶ 35 n.113). I agree that this would be nice to include. However, public, granular data on e-book sales do not exist, and, though it is my understanding that the Internet Archive requested such data, the publishers did not provide e-book data by month dating back to the time at which Plaintiffs acquired rights in the Works in Suit.[1] I would be happy to add an analysis of the effect on e-book sales if data were made available. I note that Dr. Prince did not offer analysis of e-book sales.

II. THE ANALYSIS OF THE EFFECT ON AMAZON SALES RANKS

2. Dr. Prince claims that my analyses on the effect of the Internet Archive on print book sales on Amazon are misleading. He points out three reasons why he thinks my analysis may be misleading: (1) there could be other events related to a particular book that drove

---

[1] *See* ECF No. 47 (IA's letter brief to court explaining the need for monthly data about the Works in Suit and the need for monthly data about works not at issue); *see also id.* at 3 ("Hachette, Penguin Random House, HarperCollins, and Wiley have produced annual commercial performance data for the works-in-suit. They do not contend that monthly data does not exist; they contend only that producing monthly data would be more difficult, so they have not produced it. . . . Monthly data is necessary because each book began to be available for digital lending on a particular known date, so it makes sense to compare the commercial performance *before* that particular date with the commercial performance *after* that date. Purely annual data does not provide enough detail, particularly because sales of a particular book change so drastically within a year."); *see also* ECF No. 49 (Plaintiffs' response letter brief).

demand—examples include marketing efforts (Prince Rpt. ¶¶ 83, 85 & 90) and supply chain issues (*id.* ¶ 89), and general trends in demand, such as a growing interest in racial justice (*id.* ¶¶ 87–88); (2) ranking data at Amazon do not explicitly capture sales on a particular day (*id.* ¶ 92); and (3) it is possible that the timing of availability of a book through the Internet Archive was driven by expectations of particularly high demand (*id.* ¶ 93). In the following paragraphs, I respond to these points.

3.   The first point that Dr. Prince raises is the lack of controls for other events that may have happened around the time of a title's inclusion in the Internet Archive. If these events increased the demand for a title independently of its availability at the Internet Archive, then the fact that I could not find a negative effect of availability at the Internet Archive on print sales may be hiding a combination of a negative effect of the Internet Archive and a positive effect of the other events. As set forth below, however, I do not believe that the hypothetical confounders identified by Dr. Prince affect the outcome of my analysis.

4.   The first set of possible confounders revolves around title-specific promotions and marketing efforts. Dr. Prince mentions the possibility of newly released editions (Prince Rpt. ¶ 90), promotions that are related to prices (*id.* ¶¶ 83, 90) and star ratings (*id.* ¶ 90), and movie releases (*id.* ¶ 83). Finally, he raises possible supply chain issues (*id.* ¶ 89) as a further explanation. None of these seem to explain my initial findings, that I can't identify a negative effect of availability on the Internet Archive on print unit sales:

> a.   Newly released editions are unlikely to positively affect the sales of existing editions of a title. In fact, they would be more likely to decrease sales of existing editions because they can be seen as close substitutes and could compete for (or "cannibalize") sales from existing editions. Note

that in my original report, I specifically estimated the effect of availability

at the Internet Archive on the demand for book editions, rather than titles,

so that Dr. Prince's concern about the effect of releases of new editions for

the Works in Suit on my analysis is unfounded.

b.   Dr. Prince raises concerns that demand for an edition may have changed

due to price decreases or changes in Amazon star ratings. While changes

in Amazon star ratings are unlikely to arise from any book-specific

promotions unless, perhaps, someone elicited fake reviews in order to

raise the ratings, prices may be designed to increase demand. I have

information on daily prices and star ratings, as well as the number of

reviews, so I can perform my analysis again taking this information into

account. Specifically, I re-ran the original regressions but added price,

average star ratings, and the number of reviews as control variables to the

regression equation. While the original equation examined the "effect" of

Internet Archive availability on an edition's Amazon rankings without

controlling for possible changes in prices and ratings, this new regression

(with the "control variables") explicitly holds prices and ratings constant.[2]

To illustrate the impact of adding these control variables, I examine the

effect of availability through the Internet Archive's Controlled Digital

---

[2] Mathematically, one can think of the coefficient on Internet Archive availability as a partial
derivative with respect to availability. We treat all other variables in the equation as constants
and examine what happens to rankings when we raise availability by one unit. In the original
regression, there were no other variables, which implies that I allowed prices and ratings to vary
rather than keep them fixed. Here, I include these other variables and they are therefore held
constant for the identification of the coefficient on Internet Archive availability.

Lending Program on rankings. Without these control variables (in my original report), I found a coefficient of 0.0024 (and a standard error of 0.002). This effect was not statistically distinguishable from zero, as I explained in my original report.[3] When I add the control variables, the coefficient becomes -0.0002 (with a standard error of 0.002) – a number even closer to zero, and even less distinguishable from zero.[4] That is, the addition of the control variables does not make a difference.

c.   Dr. Prince also mentions the possibility of film and TV adaptations (Prince Rpt. ¶ 83). While this would only be of concern if the timing of these adaptations coincides with the timing of changes in availability at the Internet Archive, I searched on Google for each title together with the word "adaptation" to identify books that were adapted into TV shows or movies between 2011 and 2021.[5] Removing these titles from the analysis

---

[3] While the coefficient is not distinguishable from zero, the point estimate suggests a 0.24% worsening in rankings.

[4] Taking the point estimate as given, it would suggest an improvement in rankings of 0.02%.

[5] When the initial search for a title suggested there might be a TV or movie adaptation, I searched more deeply to identify potential matches. Using this approach, I identified 19 such titles: *A Dance with Dragons; Caliban's War; Daring Greatly; Escape from Mr. Limoncello's Library; Fallen; Gone Girl; I Funny; If I Stay; Leviathan Wakes; Middle School the Worst Years of my Life; Night Watch; Simon vs. the Homo Sapiens Agenda; Station Eleven; The Extraordinary Education of Nicholas Benedict; The Innocent Man; The Miseducation of Cameron Post; The Mysterious Benedict Society; The Mysterious Benedict Society and the Perilous Journey; The One and Only Ivan*. Note that this list might not be exhaustive or may include a book title that was not adapted because a movie's title may be different from the book's title (e.g., *Simon vs. the Homo Sapiens Agenda* was adapted into a movie called *Love, Simon*; and *A Dance with Dragons* is technically a prequel to the *Game of Thrones* TV series). While my list may be imperfect, Dr. Prince does not make any attempt to identify titles associated with adaptations. If such a list were made available, I would be happy to incorporate it.

does not change my results in a meaningful way. For example, if I drop all editions of these titles and repeat the analysis from above (with price and ratings controls), I now find a coefficient of -0.0016, with a standard error of 0.002. The estimated coefficient suggests a 0.16% improvement in rankings due to the controlled digital lending program, although again this estimated effect is not statistically distinguishable from zero. Because this coefficient is very close to that in Paragraph 4b, I conclude that movie and TV adaptations of a title do not seem to drive the results.

d.   Finally, it is unclear to me why Dr. Prince's argument of supply chain issues poses a problem for my analysis. There is one pattern that could lead me to miss a negative effect of Internet Archive availability on sales by not accounting for supply chain issues: if these supply chain issues only apply to other books that compete for similar rankings as the Works in Suit, but not the Works in Suit themselves. I don't see a reason why this would be the case. Note here that my use of rankings is helpful, because many events that affect the entire industry (like supply chain issues) only pose a threat to my identification strategy if they affect the Works in Suit differently from comparable books. Again, I find this unlikely, and Dr. Prince provides no reason to think that it would affect some books differently than others.[6]

---

[6] In Paragraphs 84 and 89, Dr. Prince refers in footnotes to an article that was published in October 2021 – long after changes in Internet Archive availability took place for the Works in Suit.

5.   A second set of confounders includes possible changes in overall interest for a genre or subject. For example, in 2020 public interest in topics around racial justice and the election may have raised demand for books on these topics (Prince Rpt. ¶¶ 87–88). If the Works in Suit are disproportionately likely to cover these topics (compared to other books), and if these books became disproportionately available at the Internet Archive shortly before these jumps in interest, then it is possible that I would falsely find a positive or null sales effect of Internet Archive availability. This concern is unlikely to affect my analysis of the effects of controlled digital lending availability because there was no one particular date on which all books were added and thus the timing of Internet Archive availability relative to shifts in interest in, say, racial justice would vary across titles. Still, I can address this concern by examining different genres. In particular, the Keepa data that I described in my initial report include information on a book's genre, which allows me to distinguish between fiction and nonfiction books. Concerns about changing interest in topics seem most salient for nonfiction books: beyond TV and movie adaptations or idiosyncratic reasons for notoriety of a particular book (such as some news event relating to the book's author)[7], the interest in fiction books is unlikely to see large shocks several years after their original publication, whereas nonfiction books may discuss certain topics for which interest rises for other reasons. To address the possibility that my results are due to large changes in interest, I repeat my analysis after dropping all editions of nonfiction books.[8] Across all analyses, the main result remains robust even when I drop nonfiction books: Considering the 30 days before an event and the 60 days after the event, I find coefficients that are very close to

---

[7] ██

██████████████████████████████████

[8] I identify 295 editions as nonfiction books in my dataset. The editions and their Amazon identifiers (ASINs) are included as Attachment C to this report.

6

zero in size and statistically indistinguishable from zero for both addition to and removal from controlled digital lending.[9] The estimated effect of the National Emergency Library continues to suggest a strong improvement in rankings but is subject to the same concerns as the analyses in my initial report: the placebo analysis that chooses a placebo National Emergency Library date one year earlier shows similar trends, so while the National Emergency Library fairly clearly did not *reduce* print book sales, I cannot conclude that the National Emergency Library *increased* print book sales.

6.   Dr. Prince's second main issue around the analysis of the effect of IA availability on rankings at Amazon was that these rankings are not a reliable measure of sales (Prince Rpt. ¶ 92). Dr. Prince argues that because Amazon is vague in describing its algorithm for sales rankings, it is possible that rankings are not based on sales revenue. Note here that rankings are likely based on unit sales rather than sales revenue—in fact there is academic work that points to this relationship.[10] However, the assertion that authors can manipulate Amazon sales rankings through marketing schemes (*see* Prince Rpt. n.256) is irrelevant here unless authors of the Works in Suit (or the pertinent Plaintiffs) exerted additional effort to increase sales after the works were made available at the Internet Archive, and more so than for works written by other authors. I find this unlikely. Note that even if this were an issue, the fact that I look at rankings would not exacerbate it because the rankings are a function of unit sales.

---

[9] The estimated coefficient on inclusion in controlled digital lending is 0.015 (standard error = 0.011), and the estimated coefficient on removal from controlled digital lending is 0.009 (standard error = 0.011). Both are close to zero, and I cannot conclude that they are different from zero.

[10] *See* Chevalier, Judith, and Austan Goolsbee. "Measuring prices and price competition online: Amazon. com and BarnesandNoble. com." *Quantitative Marketing and Economics* 1.2 (2003): 203-222.

7.   I agree with Dr. Prince that for my analysis to uncover causal relationships, I have

to ensure that the decision to make a book available for digital lending was unrelated to

(expectations of) periods of particularly high demand (Prince Rpt. ¶ 93). While I have little

reason to believe that availability for digital lending was driven by anything other than the

passing of the five-year period and digitization capacity constraints at the Internet Archive, I am

adding to this discourse here by referring to depositions that showed that digitization was done

with no regard to a title's concurrent or impending demand but rather after a certain delay has

passed.[11] In fact, by relying on testimony that confirms that the Internet Archive made titles

available for digital lending solely on the basis of which titles were available, the argument for

an exogenous change in digital availability in my report is arguably stronger than many other

analyses in other, peer-reviewed papers. For example, in my 2016 paper on the effects of private

copyright protection—which Dr. Prince discusses in some detail in his report—I confirm the

exogeneity of piracy protection in two ways. First, I rely on conversations with people who

worked at the anti-piracy company that was the subject of my 2016 paper, who suggested that

the identity of titles did not play a role in determining the order of protection. Second, I used

Google Trends search volume to show that the overall title interest (measured in monthly Google

searches rather than direct demand information) is not correlated with the timing of piracy

---

[11] Mills Dep. 82:6-12 ("Q. Do you know generally how in-copyright books were selected for
digitization in sort of the early years in which Internet Archive was digitizing in-copyright
books? A. I would say availability in terms of donations to the Internet Archive for acquisitions
by the Internet Archive."); *see also id.* 71:13-72:1, 90:12-91:11.

Note also that from my own work on Google Books digitization (Nagaraj, Abhishek and
Reimers, Imke, "Digitization and the Demand for Physical Works: Evidence from the Google
Books Project," SSRN Working Paper (Apr. 12, 2021)), which Dr. Prince mentions in his
report), we also learned that no attention was paid to which books were digitized first, but rather
that focus was placed on making books available digitally as quickly as possible.

protection. The direct sworn testimony regarding the decision to make books available for digital lending here eliminates the need for that sort of analysis.

8.   A related point raised by Dr. Prince is that I did not "formally test the change in sales with season" between the period of my analysis and the placebo period (Prince Rpt. ¶ 91). Because my analysis uses (relative) rankings rather than (absolute) sales, the concern about changes in sales across seasons is alleviated. For example, I do not need to account for periods of particularly high book sales (such as Christmas) because *all* books likely see increases in demand and rankings describe sales relative to other books. Still, I can allow for the possibility that the Works in Suit are particularly popular during certain times of the year. I do this in a new analysis that further extends the original analysis. In particular, in addition to the control variables for an edition's price and ratings, I now include a control variable for each month of the year. Much like a control variable for the price allows me to look at the effect of Internet Archive availability while holding the price constant, the control variables for each month of the year allow me to examine the effect of digital lending by Internet Archive while holding fixed the month of year—in other words, to control for the seasonality of demand for the Works in Suit. Adding these additional controls, I now find that the coefficient on the Internet Archive availability through Controlled Digital Lending is -0.0007 (standard error = 0.002). Again, the estimated effect is very small (equivalent to a ranking improvement of 0.07%), and not statistically distinguishable from zero. In addition to finding no evidence that seasonality affected my estimates, I note that Dr. Prince also did not perform any analysis on the role of seasonality.

III.   THE ANALYSIS OF LIFETIME SALES

9.   Dr. Prince also comments on my analysis of the lifetime sales of the works in suit. He makes the following points: (1) one cannot talk about a book's lifetime sales while the book

is still being sold (Prince Rpt. ¶ 114); (2) the relevant margin is not that of an individual title's sales progression, but rather the share of sales of the publisher's entire catalog that can be attributed to "old" books (*id.* ¶ 110); (3) one should look at the lifetime sales of a title, starting from the publication of its first edition, rather than the lifetime sales of an edition (*id.* ¶¶ 110 n.282 & 115); and (4) it would be more appropriate to consider revenue rather than unit sales to discuss market harm (*id.* ¶ 115). In the following paragraphs, I respond to each of these points.

10. While I agree that one cannot talk about a title's lifetime sales when the title still sees positive sales today (Prince Rpt. ¶ 114), I believe my analysis of the sales decay in the first ten years is valuable. This seems like a semantic point rather than a substantive one: instead of calling it lifetime sales, I am happy to change my wording to lifetime sales *to date*. Still, I believe the analysis is informative. In fact, the analysis in my initial report can be used to create lower and upper bound estimates of the share of a book edition's sales that occur in the first five years after original publication. For example, in my initial report I examined the sales trajectory of the editions of the Works in Suit by Penguin Random House that were published in 2010, between 2010 and 2020.[12] I found that on average 90% of an edition's unit sales occurred in years 0 to 5 after its publication, and only 10% in the following five years. Note here that Dr. Prince pointed out in Paragraph 115 that I reported the sales share in years 0 to 5, i.e., the first six years, relative to the first 11 years. My representation requires additional explanation because the "first" year after publication is not easily defined. For example, for editions that were published in January, one would expect many sales in the first year. By contrast, if an edition was published in December, the "first-year" sales might better be reflected in the following

---

[12] Recall that I used editions published in 2010 to ensure that I would observe the entirety of the first five years since publication. *See* Paragraph 13 below that explains the implications of using the publication dates of editions rather than titles.

calendar year. In my original report, I used years zero to five, assuming that the first year was the year *after* publication. If I instead consider the years zero to four relative to the first 11 years, the share at Penguin Random House is 86%. A similar concern could apply to my analysis of the remaining publishers. For these, I used the more conservative measure (the first year was the year of publication, even if the edition was published in December) to begin with, and a more liberal interpretation would have significantly raised the "first-year" share. Specifically, for Hachette, the share of unit sales that occurred in years zero to one, relative to years zero to three, was 78%, rather than 65%; and for HarperCollins, the share of unit sales that occurred in years zero to one, relative to years zero to three, would be 82%, rather than 74%.

11.   Based on the numbers for the Works in Suit from Penguin Random House, we can provide bounds for the share of unit sales that occur in the first five years compared to, say, the first 50 years.[13] This is an extension to my initial analysis. The upper bound of this five-year share is 86%—this would be the case if no units were sold more than 11 years after an edition's publication. I can create a lower bound of the five-year share by assuming that unit sales remain constant after the first five years, so that annual average sales in years six through 11 are equal to average annual sales in years 12–50. If this is accurate, then the "lifetime" share (defining the lifetime as 50 years) of unit sales that occur in the first five years would be 45%.[14] Because an edition's unit sales in subsequent years are unlikely to fall below zero or rise above

---

[13] Here, to remain very conservative, I will use years zero to four to define the first five years.

[14] To calculate this share, I divided the 90% from the first five years by the sum of all "shares" from the years six – 50. Because the first five years account for 86% of the 11-year share, the next six years account for 14%. Per year, this would be 14/6 = 2.33%. Now suppose these 2.33% extend for each year from year six to 50 – i.e., 45 years. This would imply that sales from year six to 50 would be 45*2.33% = 105% of the total sales in the first 11 years. Now I can compare sales in the first five years (86% of the 11-year total) with sales in the next 45 years (105%). As a percent of the total, the first five years make up 86 / (86 + 105) = 45%.

the level of previous years, the true share of sales in the first five years is likely between 45%
and 86% for most titles.

12.   Dr. Prince also asserts that the damages to publishers are more accurately
described as a share of sales of the entire catalog, rather than on a title-by-title basis (Prince Rpt.
¶¶ 110-111). While I agree that the share of first-five-year sales relative to a publisher's entire
catalog is relevant, I did not have access to this type of information.[15] The shares of "backlist"
sales that Dr. Prince mentions in Paragraphs 52 and 111 (and which I also mentioned in my
initial report) are irrelevant here, because the publishers define the "backlist" as titles that are
more than one or two years old, not more than five years old.[16] However, I did not receive any
information—nor am I aware of publicly available information—about the share of sales from
books that are between two and five years old. Because information for the relevant time spans
on the catalog was not available, I focused on information that was available: the share of an
edition's sales that occurred in the first few years of the edition's availability for sale or license.

13.   Dr. Prince further contested that a *title's* lifetime sales trajectory is different from
an *edition's* lifetime sales trajectory, and that looking at book editions was the wrong dimension
(Prince Rpt. ¶ 115). Here, I respond to this comment in several ways:

>   a.   I chose to use edition sales merely for data availability reasons. Other
>   than Wiley, the longest time series of edition-level and title-level sales (or
>   revenue) data I had available was from 2010 (Penguin Random House). If

---

[15] *See* ECF Nos. 47, 49.

[16] Hachette defines backlist as books published two or more years ago, front list as books
published within one year, and prior year published as books published between the front list and
backlist terms. Sevier Dep. 63:4-67:3. Penguin Random House, Wiley, and HarperCollins
define backlist as books over a year old and front list as books younger than one year. Weber
Dep. 101:21-104:9; Pavese Dep. 215:21-217:9; Restivo-Alessi Dep. 56:10-13.

I want to know the share of sales that happened in the first five years after a title's initial publication, I am limited—as Dr. Prince correctly points out—to titles that were originally published in 2010. For Penguin Random House, for example, this includes two titles in the Works in Suit. Two titles are not enough to make meaningful inference about any sales trajectories more generally. Instead, I opted to follow sales of specific editions, which allowed me to increase the number of data points on which to base my findings. I understand that the Internet Archive requested data from publishers going back to the times at which Plaintiffs acquired rights in the Works in Suit and that Plaintiffs declined to produce such information.

b.   Note that my decision to look at editions is unlikely to bias the sales decay over time in favor of finding a larger decay. In fact, it seems more likely that my estimate of the sales decay over time is lower than the true decay. Whereas sales for a *title* likely peak early and then decrease rather quickly, sales of a subsequent edition—perhaps several years after the initial publication – may never experience the same spike in demand early on because those who were most eager to read the book would have already bought an earlier edition. That is, if anything, my analysis under-estimates the sales decay and therefore the share of a title's sales that occur in the first five years.

c.   To illustrate differences in the sales decay for the first editions versus later editions, one could look at the different editions for *Theodore Boone: Kid*

13

Lawyer, which Dr. Prince pointed out (Prince Rpt. ¶ 115) to be one of the two Penguin Random House titles of the Works in Suit that were originally published in 2010. The hardcover edition of *Theodore Boone* was published in May 2010, and its sales (and revenues) in 2010 and 2011 accounted for 96% of its total sales (and revenues) from 2010 to 2020. Three other editions of the title were published in 2011 (one mass market paperback and two e-books). For those, sales (revenue) in 2011 and 2012 accounted for 70% (71%) of their total sales (revenues) from 2011 to 2020.[17]

d. Because I observe sales and revenue for Penguin Random House books from 2010 to 2020, I can also compare the sales and revenue decay for the first and later editions of other titles that were originally published after 2010. In a supplementary analysis of the editions' lifetime sales, I keep titles with an original publication date after 2009 and which have subsequent editions in later years before 2016, to allow for sufficient observations to measure sales decay. I follow unit sales and revenue for all editions of these titles in their first five years. For the initial editions of these titles, I find that the average share of unit sales (revenue) in the first two years – compared to the first five years—was 75% (77%), whereas the same share was 67% (67%) for editions published in later years. This

---

[17] Note that all print and ebook editions of the other 2010 title, *Who Was Jackie Robinson*, were published in 2010, which prevents me from examining whether the sales decay for the first editions is different from the sales decay for later editions.

suggests that the original editions indeed see faster sales decay than later editions.

e. Dr. Prince also mentioned two specific titles which had relatively high sales several years after their original publication (Prince Rpt. ¶ 110): *Daring Greatly*, and *A Dance with Dragons*, sales for both were likely bolstered by releases of related TV shows. Note that the sales trajectories for these titles are likely outliers, particularly because they of their connections to television series. (This is bolstered by Dr. Prince's own suggestion, discussed above, that a movie or television adaptation of a work might lead to an increase in sales.)

14. Dr. Prince also posits that the relevant margin is on the publishers' sales revenue rather than unit sales (Prince Rpt. ¶ 115). In my initial report, I used unit sales for two reasons. First, I did not believe the choice to use unit sales as opposed to revenues would make a large difference because the prices that publishers receive per unit of a specific edition remain quite constant. Second, unit sales are a more appropriate indicator of consumer demand and whether units of digitized books serve as substitutes for either e-books or physical books. Still, in this reply report, I test my claim that the choice between unit sales and revenues does not drive my results by repeating my analysis using revenue information. For the Works in Suit at Penguin Random House, the share of unit sales in the first five years as opposed to the first 11 years was 86%. The share of revenues over the same time span was 85%. For the Works in Suit at Hachette, I had found that 65% of the five-year unit sales occur in the first year (although this number rises to 79% if I drop audio books). If I instead report the same share in terms of revenues, it is 66% if I include audiobooks and 79% if I drop them. For HarperCollins, the

equivalent share of unit sales reported in my initial report was 74%. If I instead report the share in terms of revenue, it is 72%. I summarize these shares in the table below. The differences between the trajectory of unit sales and that of revenues is very small for three of the four plaintiffs, and for the fourth (Wiley), it is moderate in magnitude.

Publisher[18]	Share definition	Unit sales	Revenue
Penguin Random House	First 5 years rel. to 11	86%	85%
Hachette	First year rel. to first 5	79%	79%
HarperCollins	First year rel. to first 4	74%	72%
Wiley	First 5 years rel. to 12	56%	46%

15.   Note here that in my initial report (¶¶ 9 & 25) I also reported the shares of sales (or revenue) of backlist titles. There, I reported the share of unit sales for HarperCollins because its spreadsheets listed unit sales rather than revenues (and for print books only).[19] For Hachette and Penguin Random House, I reported revenue shares, because their spreadsheets did not list unit sales.[20] Because my *title-level analysis* just above suggests that unit sales and revenue shares are very closely related, I do not believe that this inconsistency in the way the data was made available to me affects my conclusions.

---

[18] This analysis excludes audiobooks. *See* Prince Rpt. ¶ 43 ("My analysis and opinions focus on the sale and licensing of print and ebook formats to the library and retail markets. There are other formats, including audiobooks, and other markets, such as schools, which are outside the scope of my analysis.").

[19] See HC0030132.

[20] See HACHETTE0012377 and PRH0072194.

IV.   OTHER SCHOLARSHIP

16.   In addition to his comments about my analyses, Dr. Prince analyzes other scholarship, including some of my own.

17.   In Paragraph 32, Dr. Prince reports that most (29 out of 33) peer-reviewed studies summarized in Danaher, Smith and Telang (2020) find that sharing of unauthorized media harms legal media sales. This is consistent with what I report in my 2016 paper. However, most of the works surveyed in the paper by Danaher et al. discuss the music and movie industries, which are different from book publishing. In fact, the paper by Danaher *et al.* only reports on one study on the book publishing industry—mine. The industry setting matters. For example, Dr. Prince cites a paper by Stanley Liebowitz in his footnote 94, which suggests that different types of media are consumed differently. In a working paper (Nagaraj and Reimers (2021)), which Dr. Prince mentions in his report, I argue that the case for book publishing can be different from those for music and movies. For example, if the quality of the digitized version is lower than that of the original, then people might check out the digital version in order to decide if they want to buy the full book. In footnote 94*,* Dr. Prince mentions that consumers are more likely to purchase a product after consuming the free version if the product is likely to be consumed multiple times, such as in music as opposed to movies. Books likely fall somewhere between the two, because it usually takes much longer to read a book than it takes to watch a movie.

18.   Even within an industry, different studies find different effects of free, digital provision of works. The fact that most but not all studies find harmful effects indicates that the effects of digital sharing can vary on a case-by-case basis as well as depending on the type of media at issue. Therefore, the question of whether the conduct of the Internet Archive harms publishers cannot be answered by looking at the effects of digital distribution in settings and

17

industries that are different from the Internet Archive. For that reason, I attempted to specifically study the effect of availability at the Internet Archive.

19.   Dr. Prince also discusses a paper I published in 2016 in the Journal of Law and Economics (*see*, e.g., Prince Rpt. ¶¶ 29-30). The results in my 2016 paper could be indicative of the effects of availability at the Internet Archive if the Internet Archive's conduct was akin to that of the infringers I studied. But, based on my understanding, that is not the case. In my 2016 paper, I report that 69.8% of infringing content in my dataset is found on cyberlockers (Internet hosting services designed to host users' files) and that "most" of the infringing content is in HTML and PDF formats. Dr. Prince points out that the format in which content is displayed thus is likely to look similar to the content made available at the Internet Archive (*id.* ¶ 35).[21] However, a few things are different between the setting in my 2016 paper and the Internet Archive. First, the digital lending interface at the Internet Archive has more of an appearance of a library than the cyberlocker hosts, as it allows people to *borrow* titles for limited time periods rather than to keep them on their computers indefinitely. The fact that one does not *own* the book when reading it via Internet Archive can diminish the experience of reading it, making consumers more likely to purchase a full edition later. Second, in my 2016 paper, I cannot distinguish what type of content leads to harm because my variable of interest describes whether the company has found an infringing website for the title at all, and not which type of content was taken down. If e-book sales only increased after a certain type of content was removed, and

---

[21] Note here that Dr. Prince also mentions a working paper by Abhishek Nagaraj and me (*id.* ¶ 38), in which I find that digitization through the Google Books project can in fact increase sales of physical book editions. While I agree that the set of books in that study is sufficiently different from the Works in Suit, the format of the digital content is quite similar to that at the Internet Archive. This at least suggest that the format itself is not enough to draw a clear conclusion from my 2016 study.

18

if that content is different from that offered through the Internet Archive, then I cannot draw inferences from that study either.[22] Third, I am not able to compare the scale (and number of site visitors) of the Internet Archive to that of the infringing sites in my 2016 paper. If traffic to the infringing sites in my 2016 paper was higher than traffic to each book's page on the Internet Archive, then the Internet Archive might have been too small to cause a sales decrease. Fourth, in my 2016 paper I report that on average titles have a total of 88 infringing sites. It is likely that removing one site has little effect on sales through other channels. Fifth, the types of books that I studied in my 2016 paper may be different from the Works in Suit, including consumers' willingness to pay positive prices for them.

20.   Because all five features could lead to different effects of availability at the Internet Archive on book sales, a separate analysis on the specific effects of the Internet Archive seemed prudent. Again, I would be happy to perform a similar analysis on e-book sales if the data were made available to me.

V.   **SUMMARY**

In summary, while I agree with some of the points that Dr. Prince made, I incorporated these points in new analyses. These new analyses confirm my opinion from my initial report.

Dated: May 27, 2022

DocuSigned by:
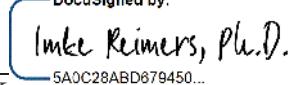
Imke Reimers, Ph.D.
5A0C28ABD679450...

IMKE REIMERS, PH.D.

---

[22] In my 2016 study, Figure 3 shows that ebook sales only increase about five months after protection begins. I did not match this timing with the types of content being taken down.